# Universal Arbitrage-free Estimation of State Price Density

Qi Hu, David Newton*

*School of Management, University of Bath, United Kingdom*

**Abstract**

Given the valuable information content of Arrow-Debreu prices, the recovery of a well behaved state price density is of considerable importance. However, this is a non-trivial task due to data limitation and the complex arbitrage-free constraints. In this paper, we develop a more effective linear programming support vector machine (SVM) estimator for state price density which incorporates no-arbitrage restrictions and bid-ask spread. This method does not depend on a particular approximation function and framework and is, therefore, universally applicable. In a parallel empirical study, we apply the method to options on the S&P 500, showing it to be accurate and smooth.

*Keywords:* State price density, Non-parametric estimation, No-arbitrage constraints, Support vector machine regression

JEL Code: G12, G17, C61

## 1. Introduction

> "The future has to be based on more a dynamic belief in how markets work and how distributions unfold. Most of risk management technology is based on looking backwards not looking forwards and I do believe that there are huge amounts of information in market prices, in particular in option market prices, about what the forward distribution of risks are, at least as gleaned by the market, and so risk management systems have to move in the direction of forward information which is contained in derivative contracts and not so much just in looking back."
>
> -Myron Scholes (2016)

Although both practitioners and academics price securities based on models that make specific assumptions about the evolution of underlying prices (such as the Black-Scholes-Merton frame-

*Corresponding author.

*Email addresses:* Q.Hu@bath.ac.uk (Qi Hu), D.P.Newton@bath.ac.uk (David Newton)

work), they have long known that these models do not always completely conform to the facts of the real world. For example, Rubinstein (1985) documented the phenomenon of the implied volatility smile before the crash of 1987, after which researchers detected pronounced deviations from previous smile shapes. The fact that the real world involves richer sources of information than than are modeled seems obvious, but only recently has the perspective shifted from an assumed underlying process towards observed market prices and implied distributions. Thanks to the availability of options data and greatly increased computational power, estimating the state price density (SPD) using a data-driven approach has gained attention.

Methods for estimating the SPD encompass parametric and non-parametric approaches (see Jackwerth (1999), Yatchew and Härdle (2006) and Figlewski (2008) for comprehensive reviews). The parametric approach specifies the SPD as a known distribution (e.g. lognormal distribution) with several unknown parameters then calibrates these by minimizing the discrepancy between the fitted and observed data[1]. The non-parametric approach interpolates the SPD between points and selects the best fit from all possible distributions using set criteria.

Although a considerable number of papers have investigated the SPD, Figlewski (2008) nevertheless argues that estimation remains an open question and none of the techniques is clearly superior. These difficulties are not surprising, since estimating the SPD poses five challenges. First, unlike theoretically modelled options with continuous strike prices, in reality options are only traded at discrete strikes. For example, strikes for the S&P 500 Index options are usually spaced \$5 apart. Second, Hentschel (2003) suggests that market options data contain noise from various sources, such as non-synchronous prices (index prices are measured fifteen minutes apart from option prices). This makes some noise-sensitive methods unattractive. Also, there is an issue with matching market prices with average bid and ask prices. Although most papers[2] use these as true option prices, in practice it is not obvious how to specify the true price since options are traded within the bid-ask spread. Third, since the SPD lies in $[0, +\infty]$, the range of observable data is insufficient to recover the information on tails. Fourth, the SPD should satisfy no-arbitrage constraints across maturities and strikes. For example, the estimated density should be positive and integrate to one. Finally, the estimation of SPD is afflicted by the "curse of differentiation". This means that the quality of the SPD estimator will be much worse

[1] Usually this can be achieved by assuming the underlying dynamics.

[2] Except Figlewski (2008) considers a weighting function to incorporate the bid-ask spread, Monnier (2013) incorporates bid-ask spread constraints and Glaser and Heider (2012)use gussian random variables from bid and ask as input call price.

than the quality of the option price estimator because differentiation has an amplifying effect on local irregularities; small irregularities in observed option prices can easily translate into serious irregularities in the SPD, including negative probabilities.

In this paper, we apply a machine learning framework to address the challenges. We first apply a modifying data filter to supress the noise from market data and then propose a more effective approach to estimating the option price via support vector regression (SVR), a method based on statistical learning theory (Vapnik (2000), Schölkopf and Smola (2002)). Rather than attempting to solve the least squares problem, we use a $\varepsilon$ insensitive loss function to incorporate bid-ask spread naturally and estimate the SPD by differentiating fitted call option prices to avoid the "curse of differentiation". Unlike neural networks that need large amounts of input and output data to properly train the network, the SVR has a well-known ability of working well in small sample cases. Our approach is a fully nonparametric and has no strong assumption on the prior distribution. It naturally accommodates the information contained in the bid-ask spread by setting an upbound predicted error, which allow traders to specify according to their own needs. It is possible to incorporate all arbitrage-free constraints into the SVR and give an arbitrage free estimator. Finally, this is a universal method of which the tensor product estimator of Fengler and Hin (2015)is a special case, using a tensor product spline kernel.

## 2. SPD and Option Prices

Although the mathematical relationship between call price function and SPD is clear, the practical implementation of SPD calculation involves three problems. First, it is difficult to find a best fitting parametric SPD estimator since this will be highly sensitive to assumptions. For example, if the dynamics of the stock price follows arithmetic Brownian motion[3], then the SPD will be inconsistent with those following a classic geometric Brownian motion. Second, without any restrictions or assumptions on variables, there are too many to consider in the nonparametric estimator. From a nonparametric statistics point of view, high dimensional regression is hardly able to achieve asymptotic consistency. Third, both consumption-based asset pricing models and no-arbitrage asset pricing models assume the no-arbitrage condition, so a set of no-arbitrage constraints is needed in the estimator.

To overcome the first problem, we change the risk neutral numeraire $\mathbb{Q}$ to the forward

[3]for example, it is reasonable to assumes the spread options whose underlying spread is positive follows the arithmetic brownian motion.

measure $\mathbb{Q}^{\mathbb{T}}$[4]. This transformation enables the estimating framework in this paper to consider a zero interest and zero dividend rate case and avoid comparing estimating methods on the ability to deal with input parameters. Further, we apply the dimension reduction method of Aït-Sahalia and Lo (1998). By transforming an option on a stock to an option on a future[5], we use the forward price $F(t,T) = S_t e^{(r-\delta)\tau}$ to represent the information of stock price $S_t$, interest rate $r$ and dividend $\delta$ then the call price function can be converted into $C(F_{t,\tau}, K, \tau, )$. Also, by assuming the homogeneity of strike and asset price, the call price function can be reformulated as $C(k,\tau)$, where $k = \frac{K}{F(t,T)}$ is called forward moneyness. Thus, the call option price is estimated in the forward-money $k$ and time to maturity $\tau$ space and the call price is changed to

$$C(k,\tau) = \frac{C(K,\tau)e^{r\tau}}{F(t,T)} \tag{1}$$

where $F(t,T)$ is forward price. $C(k,\tau)$ is called the pre-processed call price.

Finally, for ensuring that the estimating framework is independent of arbitrage, next we derive the no-arbitrage constraints for the pre-processed call price. Existing studies have reached a consensus on the necessary and sufficient conditions to guarantee that option prices are arbitrage free and two papers offer a good insight into how to derive the no-arbitrage conditions of the call price surface. From the option strategy perspective, Carr and Madan (2005) show the idea of static arbitrage, which means there is no arbitrage opportunity on the option price surface. They show that excluding the opportunities of gaining from butterfly spread, calendar spread and other conditions are sufficient to define an arbitrage-free option price surface. From the classical mathematical finance perspective, Roper (2010) helpfully summaries the no-arbitrage condition for a call price surface based on the properties of a martingale. We sketch Theorem 2.1 of Roper (2010) in Proposition 8, then simplify the conditions for a pre-processed call price.

**Proposition 1.** *Let $S_t > 0$. Define the function $C : [0,\infty) \times [0,\infty) \to \mathbb{R}$ , such that if $C(K,\tau)$ satisfies following conditions*

*(C1) (convexity in strike price K) $C(K,\tau)$is convex function in $K$ for all $\tau \geq 0$*

*(C2)(monotonicity in time to maturity $\tau$) $C(K,\tau)$is non-decreasing in $\tau$ for all $K \geq 0$*

*(C3) (The call price is limit as strike approach to infinity) $lim_{K\to\infty} C(K,\tau) = 0$ for all $\tau$*

[4]see Jarrow(1987) and Geman et al(1995) for change of numerarie method. Also, Gope and Fries (2011)called this step as normalization of call price

[5]According to Aït-Sahalia and Lo (1998), if the future and option have the same maturity, then the European option price on stock is equal to the European option on future.

*(C4) (Price bounds) for all* $K \geq 0, \tau \geq 0$

$$max(0, S - K) \leq C(K, \tau) \leq S \tag{2}$$

*(C5) has expiry value* $C(K, 0) = max(S - K)$ *for all* $K$

*Then*

*there exists a non-negative Markov martingale* $M_\tau$ *such that for all* $K, \tau \geq 0$

$$C(K, \tau) = E[(M_T - K)^+ | \mathcal{F}_0 \tag{3}$$

*that* $M$ *is a non-negative martingale*

Following the conditions of Roper (2010) and assuming that $r = \delta = 0$, we derive the no arbitrage conditions of pre-processed call prices as follows:

**Proposition 2.** *Under the pre-processed call prices framework, it is evident that (C1), (C2) and (C4) imply*

$$0 < C(k, \tau) < 1 \tag{4}$$

$$-1 \leq \frac{\partial C(k, \tau)}{\partial k} \leq 0 \tag{5}$$

$$\frac{\partial^2 C(k, \tau)}{\partial k^2} \geq 0 \tag{6}$$

*see Appendix C*

## 3. Support Vector Machine Framework

Suppose the market price data set is $\{(x_1(k, \tau), c_1(k, \tau)), ..., (x_i(k, \tau), c_i(k, \tau)) \subset \mathbb{R}^d \times \mathbb{R}\}$, where $i$ denotes the number of observations and $d$ indexed the dimension of the input space. If we consider the (vectored) call price $C(k, \tau)$, the call price function approximation problem becomes one of finding a function $f$ such that

$$C(k, \tau) = f(k, \tau) + \varepsilon \tag{7}$$

where $f(k, \tau)$ is an estimated function, and $\varepsilon$ is an error term. In practice, as long as the estimated call price is in the bid ask range[6] (or trader-specified error tolerance), we can consider

[6]Insider the [bid price, ask price] range.

the estimated price as precise. Thus, ideally a trader-specified error tolerance is preferable in the estimation framework. This intuition coincides with the key idea of a support vector machine. As shown in the left subplot in Figure 1, the estimated values are allowed have $\varepsilon$ discrepancy from the actual call price or, to put it another way, we accept the estimated call price inside the gray area.

### *3.1. Support Vector Machine*

Theoretically, the input call price $C(k,\tau)$ could be approximated by a linear combination of any continuous function. Simply, we can use any continuous unknown function to connect the known call price dots.

$$f(k,\tau) = \sum_{i=1}^{\infty} \beta_i \phi_i(k)\phi_i(\tau) + b \tag{8}$$

Where $\phi_i(.)$ is a basis function such as a spline, polynomial, sigmoid function etc. and$\beta_i$ and $b$ are associated coefficients. Motivated by statistical learning theory, the input call price $C(k,\tau)$ can be mapped into a linear feature space by a kernel function. We define the kernel function $K(k,\tau)$ as

$$K(k,\tau) = \phi_i(k)\phi_i(\tau) \tag{9}$$

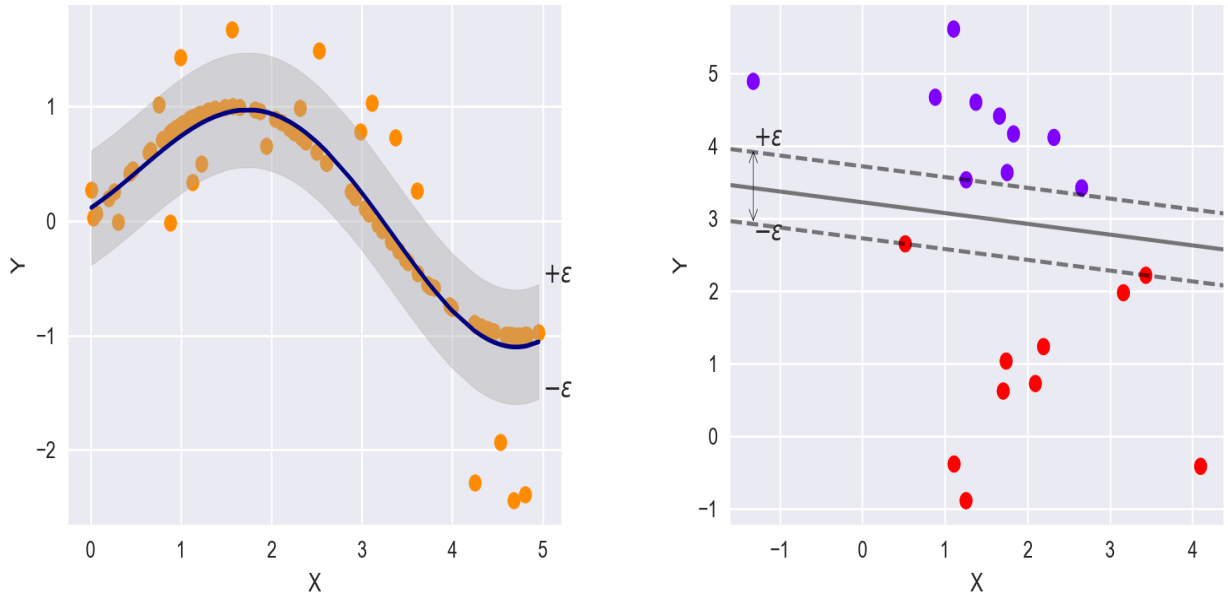
As shown in left subplot in Figure 1 and suggested by Vapnik et al. (1997), the estimated function is only influenced by the points near the dashed line. A small perturbation of data points away from the dashed line will not affect the slope and shape of the estimated function.

Consequently, we can numerically truncate the Equation (14) into

$$f(k,\tau) = \sum_{i\in SV}^{N} \alpha_i K(k,\tau) + b \tag{10}$$

where the support vector defines as input pair of $(k,\tau)$ which has non-zero associated $\alpha$ and $N$ is the number of support vectors. This remarkable feature turns an estimation problem of finding an infinity of coefficients $\beta_i$ to specifying a small number of $\alpha_i$. Moreover, from a practical point of view, we wish the call price function to be explained as simply as possible. Mathematically, this means that $f(k,\tau)$ be as flat as possible and finding the smallest '$l_0$ seminorm' of $\alpha_i$ , which has the form

**Figure 1: Intuition of Support Vector Machine**




**Note**: The right subplot illustrates the intuition of support vector machine. The left subplot presents the input space after mapping by a kernel function.

$$min||\alpha_i||_0 \tag{11}$$

However, Elad (2010) argues that a better way to address the '$l_0$ seminorm' minimization problem is to minimize the $l_1$ norm because the current algorithm to solve the '$l_0$ seminorm' problem is not efficient; thus in our support vector machine framework, our aim turns to finding

$$min||\alpha_i||_1 \tag{12}$$

subject to

$$|f(k,\tau) - C(k,\tau)| < \varepsilon \tag{13}$$

where $\varepsilon$ can be controlled and bound depending on a trader's need. Furthermore, in practice, a perfect mapping from input data to linear feature space is unobtainable since the true values contain outliers and noise; hence we apply an $\varepsilon$ insensitive loss function to penalize the deviation between estimated and true value. As we discussed before, only if a predicted value is outside

the bid ask range do we consider it as mis-priced. Formally, the loss function is

$$|\xi|_\varepsilon = \{ \begin{array}{ll} 0 & |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & otherwise \end{array} \tag{14}$$

The approximation problem is given by

$$min_{(\alpha,b)}||\alpha||_1 + C\sum_{i=1}^{N}|f(k,\tau) - C(k,\tau)| \tag{15}$$

subject to

$$-\xi \leq \alpha\mathbf{K} + b - C(k,\tau) \leq \xi \tag{16}$$

where $\boldsymbol{K}$ is a vector of kernel function and the constant $C > 0$ balances the trade-off between the flatness of the estimated model and the amount of deviation allowed. $\xi$ determines the error insensitive zone of the estimated model (the gray tube on the right panel of Figure1). If C is too large, then the objective function (21) is considered to minimize the empirical risk only and that means only caring about how well the function approximates the input data. On the other hand, if $\xi$ is too large, then we may get a flat estimated function. In our case, this could result in a noisy call price surface, which entails a multimodal state price density. Thus, the selection of C and $\xi$ is highly important and is attained via gridsearch and cross-validation techniques in machine learning theory. In this paper, we search C from $1e-3$ to $1e3$ and error tolerance $\varepsilon$ from $[0, \frac{1}{4}spread]$.

Now we reformulate the estimated problem into matrix form and replacing $C(k,\tau)$ with y, such that it becomes

$$min_{(\alpha,b,\xi,a)}\mathbf{1}^T\mathbf{a} + C\mathbf{1}^T\xi \tag{17}$$

subject to

$$\begin{aligned} -\xi &\leq \mathbf{K}\alpha + b\mathbf{1} - y \leq \xi \\ -\mathbf{a} &\leq \alpha \leq \mathbf{a} \\ 0 &\leq \mathbf{1}\varepsilon \leq \xi \end{aligned} \tag{18}$$

where $\boldsymbol{a}$ is a vector of coefficients of $||\alpha_i||_1$. The linear programing optimization program(23)

and (24) serves as our basic call price SVM estimator. Since the call price function exhibits no arbitrage, further constraints are added in the next section.

### *3.2. No arbitrage Constrained $L_1$-SVM*

As has been seen in Section 2.2, the arbitrage free pre-processed call price can be obtained through restricting the first or second order derivatives and output bound of price. In this section, we show that these constraints can be incorporated into the $L_1$-SVM framework without changing its linear programming nature.

As noted in Equation (16), the estimated call price is linear in the parameters $\alpha$ and, if we take first order derivative of the call price respect to the jth component $k^j$ of $k \in \mathbb{R}^d$, the result is

$$\frac{\partial f(k)}{\partial k^j} = \sum_i \alpha_i \frac{\partial K(\boldsymbol{k}, k_i)}{\partial k^j} = r_1(x)^T \alpha \tag{19}$$

where $r_1(x) = [\frac{\partial k(k_1)}{\partial k^j} ... \frac{\partial k(k,k_i)}{\partial k^j} ... \frac{\partial k(k,k_N)}{\partial k^j}]^T$[7]. Similarity, if we take the kth order derivative of the call price respect to the jth component $k^j$ of $k \in \mathbb{R}^d$ , the result is still linear and only depends on the kernel function and input data

$$f^k(x) = r_k(x)^T \alpha \tag{20}$$

where $r_k(x)$ is the coefficients for the kth order derivative. Let the corresponding value of call price be the vector $\boldsymbol{Y_k}$, The matrix form of monotonicity and convexity constraints $\Gamma_k(Z_k)$ is

$$\Gamma_k(Z_k) = \begin{bmatrix} r_k(x_1)^T & 0 \\ ... & ... \\ r_k(x_p)^T & 0 \\ ... & ... \\ r_k(x_{|Z_k|})^T & 0 \end{bmatrix} \tag{21}$$

[7]The form of the support vector machine implies that all derivatives of the estimated model are linear.

$$Y_k(Z_k) = \begin{bmatrix} y_1^{(k)} \\ \dots \\ y_p^{(k)} \\ \dots \\ y_{|Z_k|}^{(k)} \end{bmatrix} \tag{22}$$

where $Z_k = \{k_{1,}k_{2,\dots.}k_{|Z_k|}\}$ contains interesting points regarding derivative constraints. The choice of $Z_k$ may be vary and depend on a trader's judgement. In our case, we argue that 0 should be included as $c(0,\tau) = \frac{e^{r\tau}}{F^\tau}$. For the spline kernel, $Z_k$ is called knot[8]. Furthermore, considering the price bounds constraints as a restriction on the zero order derivative of the call price, we establish the general no arbitrage constrained SVM

$$min_{(\alpha,b,\xi,a)}\mathbf{1}^T\mathbf{a} + C\mathbf{1}^T\xi + \sum_{k=1}^{n_{sc}} \lambda 1^T z_k \tag{23}$$

subject to

$$\begin{aligned} &-\xi \leq \mathbf{K}\alpha + b\mathbf{1} - y \leq \xi \\ &-\mathbf{a} \leq \alpha \leq \mathbf{a} \\ &0 \leq \mathbf{1}\varepsilon \leq \xi \\ &\Gamma_k(Z_k)\theta - Y_k(Z_k) \leq z_k \end{aligned} \tag{24}$$

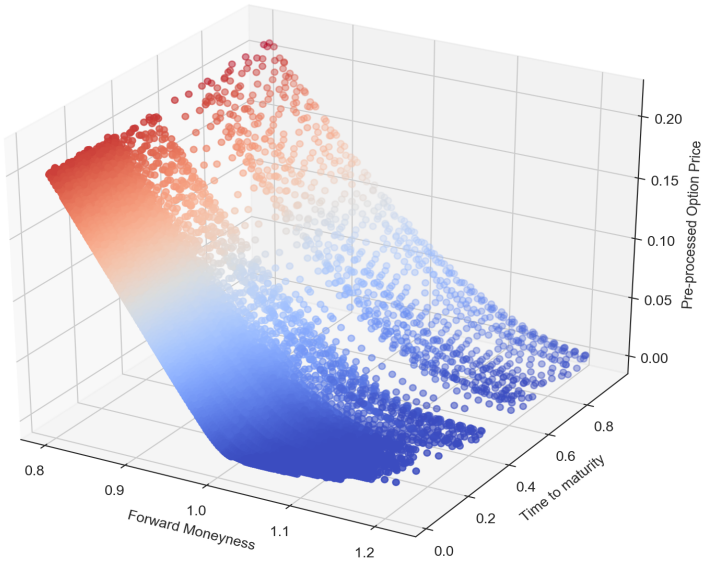
## 4. Empirical Analysis

### *4.1. Data Description*

To illustrate the effectiveness of our method, we use daily prices for S&P 500 index options from OptionMetrics. The S&P 500 index is taken as a good indicator for the U.S. market portfolio and the corresponding options are, therefore, expected to impound a consensus between market participants. The dataset includes closing index price and interest rate. As regards the option price, we collect the bid and ask price, trading volume and strike price. The dataset runs from January 5, 2000 to April 30, 2016, which yields 4,025 trading dates and 412 expiration dates. The average daily trading volume of each contract is 1321.35. Our dataset contains total 8,261,170 quotes with time to maturity varying from 7 days to 365 days. We use the average of the bid and ask price as 'true' call price but we set ε in Equation (30) as a quarter of bid and

[8]Please referred Fengler and Hin (2015)for choosing ideal knots using Akaike Information Criterion(AIC).

**Figure 2: Scatter Plot of Pre-processed Call Option Price**




**Note:** This figure plots the pre-processed call price from 01/01/2016 to 31/04/2016. The x-axis is forward moneyness and y-axis is time to maturity.

ask spread to consider the information within the bid and ask range.

Unlike a parametric method, our non-parametric $L_1$-SVM estimator is data-driven and requires arbitrage-free input data. Therefore, our dataset poses three challenges. First, the option prices are imprecise because tick sizes, bid-ask spread, and non-synchronicity of index and option prices constitute a source of error (Hentschel (2003)). The true trade price is not always centered between the bid and ask prices. Second, there are no observable data for the daily dividend yield. Third, in the money (ITM) options are less liquid than out of the money (OTM) options, with a potential impact of differential liquidity on prices.

To ensure that we have reliable option quotes and solve these challenges, we apply three

**Table 1: Descriptive Statistics of S&P 500 OTM Options Data**

Panel A: Descriptive Statistics

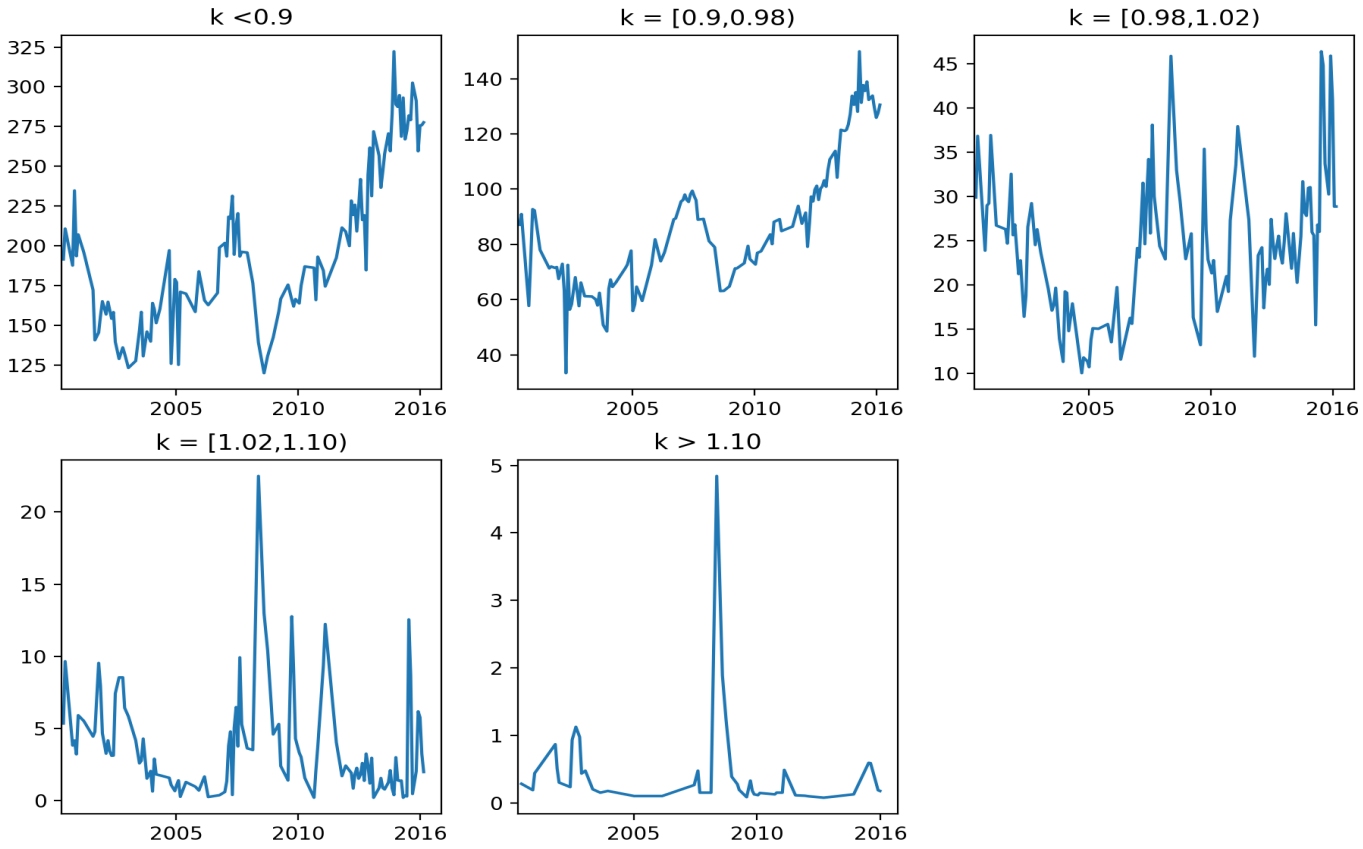
| | | | | Percentiles | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Variables | Mean | Std. Dev. | Min | 5% | 10% | 50% | 90% | 95% | Max |
| Call Price ($) | 13.31 | 15.13 | 0.08 | 0.28 | 0.50 | 8.15 | 33.90 | 43.75 | 146.3 |
| Put Price ($) | 12.77 | 14.31 | 0.08 | 0.50 | 0.93 | 7.60 | 32.20 | 42.05 | 151.50 |
| Implied Volatility $\sigma$ (%) | 20.62 | 8.39 | 6.46 | 10.30 | 11.46 | 19.19 | 31.10 | 35.69 | 87.11 |
| Strike price($K$) | 1537.75 | 377.62 | 550 | 930 | 1030 | 1540 | 2030 | 2110 | 2500 |
| Days($\tau$) | 53.82 | 41.47 | 7 | 14 | 18 | 44 | 95 | 127 | 365 |
| Index Price ($S$) | 1601.16 | 389.84 | 676.53 | 944.89 | 1091.60 | 1606.28 | 2079.51 | 2098.48 | 2130.82 |
| Trading Volume ($V$) | 1321.35 | 3690.84 | 1 | 2 | 4 | 110 | 3581 | 6698 | 157542 |

Panel B. Options by Forward Moneyness-Maturity

| Forward Moneyness $K/F$ | DTM | ITM | ATM | OTM | DOTM |
|---|---|---|---|---|---|
| | $<$0.90 | [0.90, 0.98) | [0.98, 1.02) | [1.02, 1.10) | $>$1.10 |
| | Average Call Prices ($) | | | | |
| Short-term $\tau \in (7, 60]$ | 237.04 | 107.40 | 30.30 | 6.60 | 2.05 |
| Medium-term $\tau \in (60, 180]$ | 243.74 | 120.17 | 50.02 | 15.35 | 4.60 |
| Long-term $\tau \in (180, 365]$ | 249.55 | 144.49 | 85.76 | 42.38 | 12.96 |
| | Average Implied Volatility $\sigma$(%) | | | | |
| Short-term $\tau \in (7, 60]$ | 29.96 | 20.92 | 16.05 | 15.54 | 22.36 |
| Medium-term $\tau \in (60, 180]$ | 25.26 | 19.93 | 16.75 | 14.51 | 16.82 |
| Long-term $\tau \in (180, 365]$ | 22.59 | 19.46 | 17.51 | 15.70 | 14.39 |

**Note:** Table 1 summaries descriptive statistics of our S&P 500 option dataset (total 390,320 observations). The sample period is from January 5, 2000 to April 30, 2016. The call price is calculated as average of bid and ask price. The OTM put prices are translated into ITM call prices by put and call parity. Std.Dev denotes the standard deviation from call option price.

**Figure 3: Monthly Average of 1-month Call Option Prices over Different Forward-moneyness**




**Note**: This Figure plots the time series of call price between 2000 and 2016 across different forward moneyness groups. k is forward-moneyness.

levels of the filter to eliminate the influence of liquidity and errors. We first remove identical observations from the OptionMetrics database then select liquidity data using certain criteria. Finally, we delete outliers based on the value of implied volatility and implied interest rate. To solve the second challenge and increase data quality, we linearly interpolate the interest rate and use the put-call parity relationship to derive the implied dividend (see Appendix A for details). Over the sample period, the average risk-free rate and implied dividends are 1.17% and 2.53%, respectively. After calculating the implied continuous dividend, we compute the Black-Scholes implied volatility using the bisection approach. Finally, we obtain 390,320 observations[9]. The resulting processed call prices are free of assumptions on interest rate and dividend yield.

Figure 2 shows the pre-processed call price of final dataset from 01/01/2016 to 31/04/2016. Clearly, our final dataset yields a well-defined call option price surface. Panel A in Table 1 presents descriptive statistics for our sample of S&P 500 daily prices and Panel B reports statistics for call prices and implied volatility by time to maturity and forward-moneyness groups. As expected, implied volatility shows a clear smile pattern and call price decreases as forward moneyness increases. Figure 3 compares the behavior of the monthly average of 1 month S&P 500 option prices over different forward moneyness. Consistent with Barletta et al. (2017), the ITM options exhibit non-stationary behavior whereas the OTM options are stationary.

### *4.2. Empirical Results*

Using the pre-processed S&P 500 option prices, we first compare the proposed $L_1$-SVM method using a radial basis function (RBF) and a spline kernel. The full process is summarized in Algorithm 1. As noted in Section 3.1, the algorithm's performance is determined by the parameter choice for the kernel function. We apply a grid search with 5-fold cross validation[10]to find the optimal parameters that minimize Equation(30). Table 3 reports the mean relative distance of specific kernel functions[11]. We test the univariate and bivariate approximation using RBF and cubic spline kernel in 3 sub-periods (the period before the financial crisis, financial crisis period and the period after the financial crisis). In each case, we report the relative distance of $L_1$-SVM.

We expect intuitively that the relative distance for the financial crisis period is higher because

[9]This is a reasonable number compares with Chiang et al. (2016) who get total 404,822 obervations using the same filters on S&P 500 options between 1996 and 2011.

[10]We use the Python library GridSearchCV and Cvxopt to program the code and the laptop conducts the codes is an Intel Core i7 and 2.9 GHz.

[11]The reason why we use relative distance is addressed in Section 5.

**Algorithm 1:** Approximation call price surface use support vector machine

(1) Initial ;

**Input:** Observed forward moneyness $\kappa_i$, i = 1,...N
Observed maturity $\tau_i$, i = 1,...N
Observed European call price $c_i(\kappa_i, \tau_i)$

a. Applying the three levels data filters

b. Transform option prices $c_i(K_i, \tau_i)$ to pre-processed call prices(under forward measure) $c_i(\kappa_i, \tau_i)$

c. Estimate and compare call price under forward measure

d. Transform the estimated call price under risk neutral measure Q

e. Estimate the state price density

**Output:** Estimated call price $c_i(\hat{\kappa}, \hat{\tau})$

Estimate the call price under forward measure $c_i(\hat{\kappa}, \hat{\tau})$

(1) Randomly split data D into 5 "folds" of equal size: $D_1, D_2....D_5$

(2) For i = 1,....5,

$min_(\alpha, b, \xi, a)$ $1^T\alpha + C1^T\xi + \sum \lambda 1^T z_k$
subject to Equation(30)

fit above model use defined parameters set of C, $\varepsilon$ and $\gamma$ (for different kernel parameters)

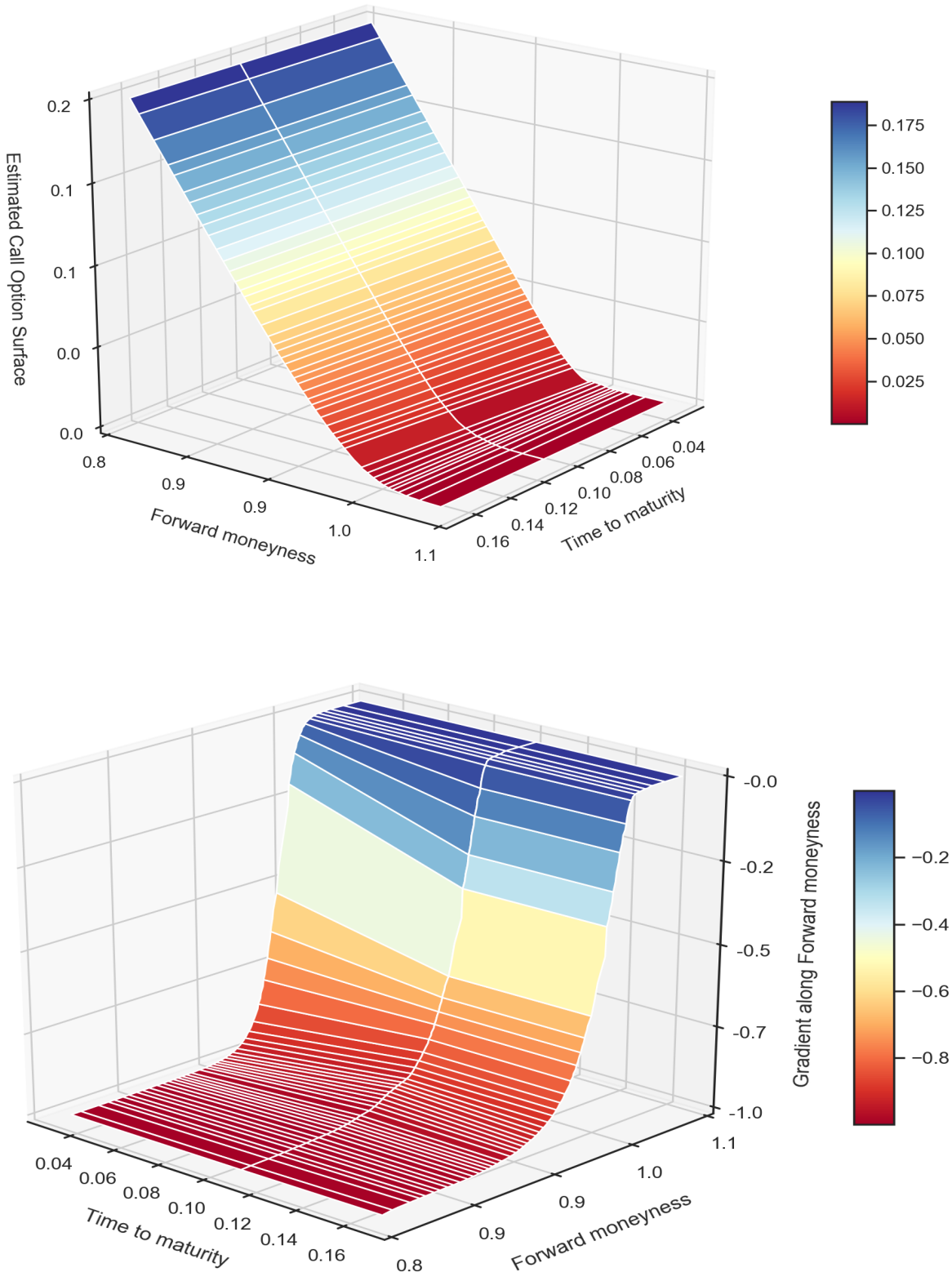
(3) Return optimal parameters of C, $\varepsilon$ and $\gamma$, which lead the minimal test error

(4) Use optimal parameters to fit the call option price function

the financial market was highly volatile in this period. Therefore, the $L_1$-SVM needs additional effort in order to fit these observations. However, as shown in Table 3, when comparing the relative distance of the $L_1$-SVM across different periods, the relative distance turns out to be decreased in the financial crisis period for both kernels in univariate and bivariate cases. This suggests that our $L_1$-SVM method is not affected by volatility in the market. The smallest relative distance is obtained by using the univariate cubic spline kernel. Also, consistent with Fengler and Hin (2015), we find that the tensor product kernel (or called surface estimator in Fengler and Hin (2015)) underperforms the univariate kernel (a slice by slice approach). In our result, the relative distances for the bivariate kernel approximation are 2.732 and 2.359 for RBF and cubic spline in the period after the financial crisis respectively, which is significantly higher than the corresponding univariate case. Fengler and Hin (2015) argue that without considering calendar-spread arbitrage, the surface estimator does not improve the fitting quality. In other words, without using calendar-spread constraints, using a surface estimator only increases the fitting difficulty since it uses a base surface to fit the call option price surface.

Figure 4 displays an example of the estimated arbitrage free call option price and its first order derivative. Consistent with Equation(10) and (11), the call option price under the forward measure is greater than 0 and less than 1. Its first-order derivative monotonically increases with

**Figure 4: Estimated Call Option Price and Its First-order Derivative**




**Note:** Figure 4 plots the estimated call price and its first-order derivative on 02/07/2013. The index price is 1614.08. The top panel shows the estimated call price. The bottom panel displays the first-order derivative of estimated call price respect to forward-moneyness.

**Table 3: Empirical Results with RBF and Spline Kernel**

| Period | Univariate | | Bivariate | |
|---|---|---|---|---|
| | RBF | Cubic Spline | RBF | Cubic Spline |
| 2000-2007 | 4.664 | 2.894 | 5.114 | 3.086 |
| 2008-2009 | 3.934 | 1.737 | 3.945 | 2.645 |
| 2010-2016 | 2.805 | 1.032 | 2.732 | 2.359 |

**Note**: This table compares the performance of $L_1$-SVM of radial basis function(RBF) and cubic spline kernel. For each kernel, we compare the relative distance of the univariate and bivariate case. Univariate refers as using a kernel function of forward-moneyness to fit each maturity slice. Bivariate refers as using a tensor product kernel function of forward-moneyness and maturity to the pre-processed call price surface.

forward-moneyness. If we reverse the change of measure (Equation(6), this first-order derivative under the risk neutral density is called delta, which measures the sensitivity of option price to change in the underlying value.

## 5. Comparison of nonparametric methods

In earlier sections, we tested the proposed method using different kernels and showed that the cubic spline kernel yields the smallest relative distance. To assess our machine learning framework and compare it with other nonparametric methods, we first summarize the differences between these nonparametric methods in two aspects:

- Assumptions of input variables: as shown in Section 2.2, the call price function can be expressed as $C(S_t, K, \tau, r, \delta)$, in which the variables $S_t$,$K$ and $\tau$ can be readily obtained from the market while $r$ and $\delta$ are difficult to calibrate. Without using the forward measure as in this paper, previous studies have different assumptions regarding these two variables (such as Glaser and Heider (2012) who assume they are both constant).

- Choice of approximation method includes two important parts: choice of kernel function and kernel dimension.

  - Various kernel functions have been applied to approximate the call option price surface in previous studies, such as low order polynomial (Kundu et al. (2016)), radial basis function (Lai (2011)) and cubic spline kernel (Fengler (2009)). In this paper, since

**Table 4: Summary of Compared Models**

| Model | Interest rate | Dividend | Kernel | Scope |
|---|---|---|---|---|
| Fengler(2009) | Deterministic | Deterministic | Cubic Spline | Maturity Slice (Univariate) |
| Fengler and Hin (2013) | Deterministic | Deterministic | Tensor product B-spline | Global (Bivariate) |

**Note:** this table compares two nonparametric estimation methods for asumption of interest rate and dividend, choice of kernel in our estimation framework.

the smallest relative distance is obtained by using a cubic spline kernel, we compare our method with Fengler (2009).

– In terms of kernel dimension, Fengler and Hin (2015) is the only study that uses the bivariate kernel approximation. In contrast with this paper, they incorporate the no-arbitrage constraints in the control net of the tensor product B-spline and solve via quadratic programming to fit the call option price surface.

To compare our $L_1$-SVM method with Fengler (2009) and Fengler and Hin (2015), we first summarize the differences between two methods in Table 4. Since our estimated framework is independent of interest rate and dividend yield, the performances of these three models are mainly determined by their optimization procedures. We review the key optimization function of Fengler (2009) and Fengler and Hin (2015) as follows.

- Maturity Slice by Slice ( Fengler (2009) )

Dividing the call option price surface into several maturity slices, Fengler (2009) proposes a method that estimates the call option price using a cubic spline function. To make the call option price surface as smooth as possible, Fengler (2009) first develops a smooth technique in implied volatility space. After obtaining the smoothed data, he transforms them back to the option price space. Using the no-arbitrage conditions on the option price, he simplifies the estimation problem as follows:

For each maturity

$$min_{g_\tau} \sum_{m=1}^{M} w_m[c(k_m,\tau) - g(k_m)]^2 + \lambda \int_{k_1}^{k_m} (g''_\tau(k))^2 dx \tag{25}$$

where $g(k_m)$ is spline series, $c(k_m, \tau)$ is the observed call price. $g''(k)$ is a regularization term introduced by Green and Silverman (1994). The optimization problem Equation (31) is solved with respect to the no-arbitrage conditions.

- Two dimensional Tensor Product Kernel (Fengler and Hin (2015))

Fengler and Hin (2015) extends the earlier method using a two univariate spline kernel. This method represents the call option price surface as a linear combination of tensor product B-spline surface. By minimizing a penalized least squares, the estimation problem becomes

$$min_\theta \frac{1}{N} \sum_{i=1}^{N} (c(\kappa_i, \tau_i) - s(\kappa_i, \tau_i))^2 + \lambda_N |\theta|^2 \tag{26}$$

where $s(k_i, \tau_i)$ is the tensor product spline, $c(k_i, \tau_i)$ is the observed call price. $\theta$ is vector of the tensor product B-spline coefficients. Without directly considering no-arbitrage in the quadratic programming framework, Fengler and Hin (2015) establish no-arbitrage conditions on an artificial kernel surface (which is called the control net of the B-spline). This approach involves a complex knot search and relocate process, which is highly time consuming.

### *5.1. Performance Measures*

When comparing the proposed $L_1$-SVM method with other nonparametric methods (Fengler (2009) and Fengler and Hin (2015)), we assess three aspects: accuracy of estimation and smoothness of call price surface with respect to forward-moneyness and time to maturity. Although the mean squared error (MSE) is a general indicator of accuracy, we use relative distance to measure the goodness of fit. This is because the total estimated number of our $L_1$-SVM method is different from (Fengler (2009) and Fengler and Hin (2015). As proposed in Section 3.1, we use a 5 fold cross-validation process to determine the optimal parameter and calculate the error, so the final estimated number equals the number of its training subset. Put simply, if we choose 5 fold cross-validation, the number of total estimated option price only accounts for 20% of the whole data. Considering this size effect, we use a relative error to measure the goodness of fit.

- **Accuracy**

We define the error as relative distance

$$RelativeDistance = \sqrt{\sum_{n=1}^{N} (\frac{C(k,\tau) - \hat{C}(k,\tau)}{C(k,\tau)})^2} \tag{27}$$

**Table 5: Empirical Results for Estimating Call Option Price**

| | | Fengler (2009) | $L_1$-SVM | Fengler (2015) |
|---|---|---|---|---|
| Relative Distance | Min | 0.314 | 0.312 | 1.164 |
| | Mean | 2.368 | 2.363 | 18.734 |
| | Max | 10.132 | 10.135 | 367.530 |
| | Std | 1.933 | 1.937 | 27.119 |
| Smoothness(Moneyness) | Min | 0.060 | 0.004 | 0.082 |
| | Mean | 35.250 | 33.369 | 43.633 |
| | Max | 4406.122 | 2854.166 | 1197.856 |
| | Std | 171.411 | 184.698 | 97.831 |
| Smoothness(TTM) | Min | 0.000 | 0.000 | 0.000 |
| | Mean | 0.006 | 0.006 | 0.010 |
| | Max | 0.116 | 0.117 | 0.231 |
| | Std | 0.011 | 0.016 | 0.020 |

**Note:** this table reports the performance of three compared methods. Relative distance and smoothness (forward-moneyness and maturity direction) is defined in Section 5. Std represents standard deviation. We use Python's Cvxopt libaray to solve the optimaztion and the laptop conducts the codes is an Intel Core i7 and 2.9 GHz.

where $C(k,\tau)$ is pre-processed call price from market and $\hat{C}(k,\tau)$ estimated by different methods. $N$ is the number of total estimated option prices.

- **Smoothness**

We use the absolute value of the second order derivative for both dimensions to measure the smoothness of the surface. We use a numerical differentiation approach, applying the central finite difference approximation.

$$\frac{\partial^2 \hat{C}(k,\tau)}{\partial k^2} \approx \frac{C(k_{m+1},\tau) - 2C(k_m,\tau) + C(k_{m-1},\tau)}{(k_m - k_{m-1})(k_{m+1} - k_m)} \tag{28}$$

$$\frac{\partial^2 \hat{C}(k,\tau)}{\partial \tau^2} \approx \frac{C(k,\tau_{m+1}) - 2C(k,\tau_m) + C(k,\tau_{m-1})}{(\tau_m - \tau_{m-1})(\tau_{m+1} - \tau_m)} \tag{29}$$

where $1 < m < N$. The smaller the second order derivatives, the smoother the call option price surface.

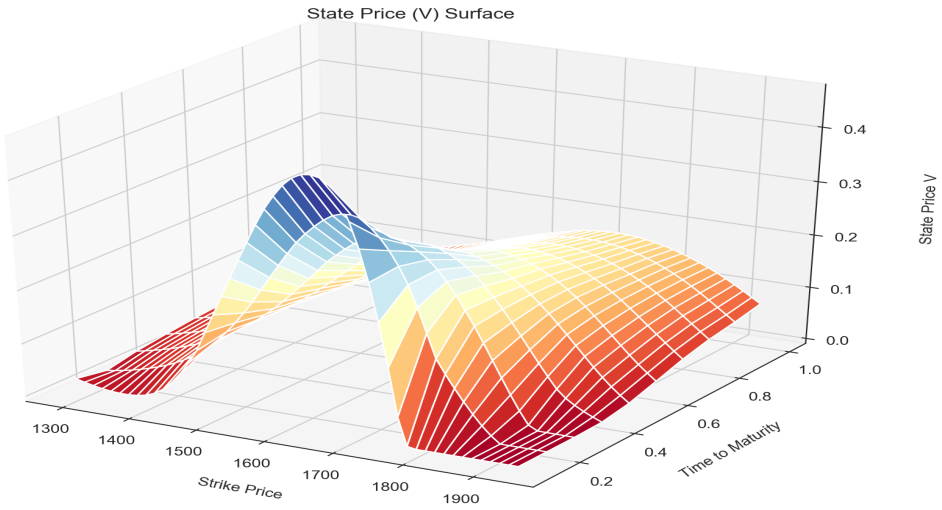
*5.2. Comparison Results and State Price Density*

Table 5 reports comparisons between these three methods. Not surprisingly, our $L_1$-SVM method shows similar results to those of Fengler (2009) because they both apply a univariate cubic spline kernel. Previous studies that use different kernel and optimization procedures can be replicated by our method. As shown in the table, compared to Fengler (2009), although our $L_1$-SVM method is slower, it shows somewhat better accuracy. This can be attributed to our comprehensive data filter approach, simple linear programming framework, incorporating the bid-ask spread information or all of them.

For relative distance, as shown in Table 5, the mean of relative distances for Fengler and Hin (2015), Fengler (2009) and our $L_1$-SVM method is 18.734, 2.368 and 2.363 respectively. Our $L_1$-SVM method produces the smallest relative distance and Fengler and Hin (2015) displays the worst performance with the mean of relative distance almost 8 times higher. When comparing the smoothness of call price surface in the forward-moneyness direction, our $L_1$-SVM method stands out in terms of stable results. Another interesting comparison among the three methods is to compare the surface smoothness in the time to maturity direction: it can be seen in Table 5 that there are no significant differences in their ability to interpolate across time. Excepting the max value and standard deviation, our method shows the same result as Fengler (2009). This finding provides empirical evidence for Fengler and Hin (2015)'s simulation result, which shows that the calendar spread no-arbitrage condition is a weak constraint. Overall, our method displays a relatively smooth call price surface.

In summary, our $L_1$-SVM method provides a universal framework that incorporates previous studies by using different kernel function in Equation(14). Previous studies with univariate and bivariate kernels can be replicated in our linear programming framework. Although our method requires additional time to search the optimal parameters, as shown in Table 5, it improves the estimation accuracy and surface smoothness in the forward-moneyness direction. Therefore, in this paper, we use the estimated call option price from $L_1$-SVM to extract the state price density (SPD). Since the above comparison is under the forward measure, based on Equation(6), we first transform the estimated option price back to the risk neutral measure and then calculate the SPD based on Equation(5). Figure 5 shows the extracted SPD on 02 July, 2013. As expected, the SPD is unimodal, smooth and positive. The SPD becomes small for far OTM and far ITM options.

**Figure 5: State Price Density on July 2, 2013**




**Note:** Figure 5 plots the state price surface on 02/07/2013. The index price is 1614.08. The state price density obtained as second oder derivatives of call price function respect to strike price.

## 6. Conclusion

In this paper, we investigate the problem of estimating risk neutral information (SPD or RND) from option prices. We find that estimation of SPD from option prices faces five challenges: (a) in Breeden and Litzenberger (1978)'s estimation equation, the strike price is continuous while strike prices in real markets are discrete; (b) market data contain noise that may lead to a coarse and multimodal SPD; (c) theoretically, the SPD starts from 0 and extends to infinity while market option data can only estimate the SPD within certain bounds; (d) the estimated call option price surface should incorporate no-arbitrage constraints; (e) the estimation of SPD suffers the 'curse of differentiation'.

A review of existing parametric and non-parametric methods shows that none of the existing methods is superior and none has successfully solved all five challenges. Specifically, the parametric method is not flexible to satisfying all no-arbitrage constraints and thus leads to under-fitting the real market data. The non-parametric method shows good performance in approximating the surface but is sensitive to the pre-determined parameters. Therefore, we propose a new machine learning approach to estimate the call option price surface. Compared with parametric and non-parametric methods, machine learning has two advantages. First, since it is a data-driven approach, machine learning exhibits good performance in solving constrained optimization problems. Second, it is not sensitive to pre-determined parameters because the

optimal values of these parameters are chosen using the gridsearch technique during training.

Although most empirical studies use the average of the bid and ask price as fair option price, the true transaction option price lies in a range. To take this into account, we develop a data-driven approach $L_1$-SVM based on standard support vector machine(SVM), which incorporates the information in the bid-ask spread in pre-defining error tolerance in a loss function. As shown in Section 3.2, our $L_1$-SVM method is sufficiently flexible to consider different models and all arbitrage-free constraints.

Empirically comparing our $L_1$-SVM method with other non-parametric methods using S&P 500 index options, we show that it is accurate and smooth, easy to implement and universally applicable through choosing different kernel functions. Previous studies that use cubic spline, low-order polynomial and tensor product spline estimation method can all be replicated within our framework.

**References**


Aït-Sahalia, Y., Duarte, J., 2003. Nonparametric option pricing under shape restrictions. Journal of Econometrics 116 (1-2), 9–47.

Aït-Sahalia, Y., Lo, A. W., 1998. Nonparametric estimation of state-price densities implicit in financial asset prices. Journal of Finance 53 (2), 499–547.

Barletta, A., Santucci de Magistris, P., Sloth, D., 2017. It Only Takes a Few Moments to Hedge. SSRN Electronic Journal.

Breeden, D. T., Litzenberger, R. H., 1978. Prices of State-Contingent Claims Implicit in Option Prices. The Journal of Business 51 (4), 621.

Carr, P., Madan, D. B., 2005. A note on sufficient conditions for no arbitrage. Finance Research Letters 2 (3), 125–130.

Chiang, M.-H., Chiu, H.-Y., Chou, R. K., 2016. Measuring the disposition effect on the option market: New evidence.

Constantinides, G. M., Jackwerth, J. C., Savov, A., 2013. The Puzzle of Index Option Returns. Review of Asset Pricing Studies 3 (2), 229–257.

Elad, M. M., 2010. Sparse and redundant representations : from theory to applications in signal and image processing. Springer.

Fengler, M. R., 2009. Arbitrage-free smoothing of the implied volatility surface. Quantitative Finance 9 (4), 417–428.

Fengler, M. R., Hin, L. Y., 2015. Semi-nonparametric estimation of the call-option price surface under strike and time-to-expiry no-arbitrage constraints. Journal of Econometrics 184 (2), 242–261.

Figlewski, S., 2008. Estimating the Implied Risk-Neutral Density for the US Market Portfolio. In: Volatility and Time Series Econometrics: Essays in Honor of Robert Engle. Oxford University Press, Oxford,UK, pp. 323–354.

Glaser, J., Heider, P., 2012. Arbitrage-free approximation of call price surfaces and input data risk. Quantitative Finance 12 (1), 61–73.

Gope, P., Fries, C. P., 2011. Arbitrage-free Asset Class Independent Volatility Surface Interpolation on Probability Space using Normed Call Prices. SSRN Electronic Journal, 1–27.

Green, P. J., Silverman, B. W., 1994. Nonparametric regression and generalized linear models. Number 58 in Monographs on Statistics and Applied Probability. Nonparametric regression and generalized linear models (58).

Hentschel, L., 2003. Errors in Implied Volatility Estimation. The Journal of Financial and Quantitative Analysis 38 (4), 779–810.

Jackwerth, J. C., 1999. Option-Implied Risk-Neutral Distributions and Implied Binomial Trees. The Journal of Derivatives 7 (1999), 66–82.

Jensen, J. L. W. V., 1906. Sur les fonctions convexes et les inegalites entre les valeurs moyennes. Acta Mathematica 30 (1), 175–193.

Kundu, A., Kumar, S., Kumar Tomar, N., Kumar Gupta, S., 2016. Call option price function in Bernstein polynomial basis with no-arbitrage inequality constraints. Journal of Inequalities and Applications 2016 (1), 153.

Lai, W.-N., 2011. Comparison of methods to estimate option implied risk-neutral densities. Quantitative Finance (December 2014), 1–17.

Monnier, J.-B., 2013. Risk-neutral density recovery via spectral analysis. SIAM Journal on Financial Mathematics 4 (1), 650–667.

Myron Scholes, 2016. Black-Scholes and beyond. Interview with Myron Scholes - YouTube. URL `https://www.youtube.com/watch?v=2Axx5f2WgHc`

Roper, M., 2010. Arbitrage Free Implied Volatility Surfaces. Working paper, School of Mathematics and Statistics, University of Sydney, NSW, Australia.

Rubinstein, M., 1985. Nonparametric tests of alternative option pricing models using all reported trades and quotes on the 30 most active CBOE option classes from August 23, 1976 through August 31, 1978. The Journal of Finance 40 (2), 455–480.

Schölkopf, B., Smola, A. J., 2002. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

Vapnik, V., Golowich, S. E., Smola, A., 1997. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. In: Advances in Neural Information Processing Systems 9. pp. 281–287.

Vapnik, V. N., 2000. The Nature of Statistical Learning Theory. Springer New York, New York, NY.

Yatchew, A., Härdle, W., 2006. Nonparametric state price density estimation using constrained least squares and the bootstrap. Journal of Econometrics 133 (2), 579–599.

Zhang, J. E., Xiang, Y., 2008. The implied volatility smirk. Quantitative Finance 8 (3), 263–284.

## AppendixA. Data Filters

Option price selection in this paper follows three principles: representative, accurate and arbitrage-free. To infer the general properties of call option prices, a representative dataset should be estimated. We construct our dataset with daily European options written on the S&P 500 index as an economically representative market portfolio. We use only the out of the money (OTM) options because these have higher liquidity and therefore the quoted prices are closer to theoretical prices. An OTM option is defined as $K < F_0$[12] for put option and $K > F_0$ for call option. To increase data quality and assuming that put-call parity holds

$$C - P = e^{-\tau t}(F - K) \tag{A.1}$$

Taking advantage of this relationship, we need to get the daily interest rate as close as possible to market to transform the OTM put prices to ITM (in the money) call prices. There are two ways to deal with the issue.

1. Interpolation

Interest rate is intuitively determined by the interest rate yield curve; therefore, we can linear interpolate the curve to obtain the daily interest rate.

2. Simple Regression

If we use the put call parity relationship and consider a linear regression of at least 4 put-call option pairs, the put call parity of Equation (4.1) can be expressed as:

$$C_i - P_i = \alpha + \beta K_i \tag{A.2}$$

where $\alpha = Fe^{-r\tau}$, $\beta = e^{-rt}$ and interest rate $r = \frac{-ln(\beta)}{\tau}$, dividend yield is $\frac{-ln(\frac{\alpha}{F})}{\tau}$. This approach enables us to back out the daily interest rate and dividend at the same time.

To ensure our dataset is close to the market, we adopt both approaches. We first linearly interpolate the yield curve to get daily interest rate using zero coupon data from the OptionMetrics database. We use this as our interest rate. Subsequently, we use Equation (A.2) to compute the daily dividend. Last but not least, we modify the data filters of Constantinides et al. (2013) and apply three levels of filter designed to obtain accurate and arbitrage-free S&P 500 option prices.

[12] $F_0$ is at the money(ATM) forward

**1. Data Accuracy Filter**

- Identical observations: we drop all duplicated observations. When observations have the same identical terms (start date, expiration date, strike, option type ) but differ in price, we keep the quotes whose implied volatility is significantly away from its moneyness neighbors.

**2. Liquidity Filter**

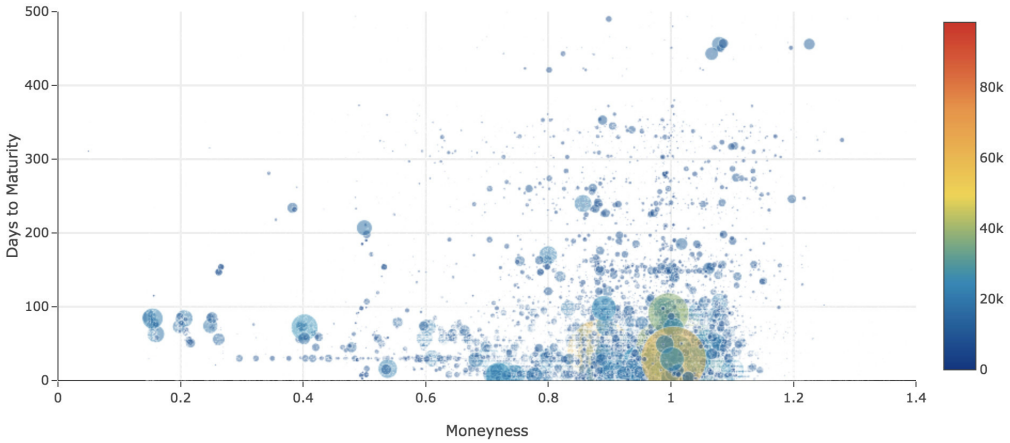
- Zero bid price : we remove quotes with zero bid prices to avoid illiquid options.
- Zero volume: in the same spirit, we remove quotes with zero volume.
- Days to maturity <7 or >365 : we remove data with maturity less than 7 days or more than 1 year. According to Constantinides et al. (2013), quotes with shorter maturity will tend to move erratically and quotes with longer maturity lack volume. We modify this filter with longer durations[13] because, from Figure A.6, S&P 500 options are still active before 1 year.
- Moneyness<0.8 or >1.2: Constantinides et al. (2013) note that option quotes in this range are thinly traded. As shown in Figure A.6, we also observe this feature in our dataset. Therefore, we remove option quotes with moneyness (the ratio of strike price to index price) below 0.8 or above 1.2.
- Implied volatility <5% or >100%: We remove all quotes with implied volatility lower than 5% or higher than 100% because as suggested by Constantinides et al. (2013), option quotes in this range can be considered as illiquid.

**3. No arbitrage Filter**

- Negative Implied interest rate: we remove quotes with negative put-call parity implied interest rate. For each available date and maturity, we use put call pairs with at least six strike prices to calculate implied interest rate[14].
- Negative Time Value: we discard quotes that have negative time value. Option prices consist of two components: intrinsic value and time value. Option quotes with negative time value show little information about investors' expectations because time value representing the amount of risk premium that investors are willing to pay.

[13]Constantinides et al.(2013) choose 0.5 years as days to maturity upper bound.
[14]We use put call parity to get implied interest that minimize implied forward pricing errors.

**Figure A.6: Trading Volume of S&P 500 on 2016**




**Note**: This Figure shows the trading volume of S&P 500 options. The data period is from 01 January 2016 to 30 April 2016.

- IV filter: we remove quotes whose implied volatility is one standard deviation away from the average among peers. We define peer group by the different levels of moneyness. More precisely, for each date and maturity, we fit the log implied volatiles in entire sample via a quadratic curve (separately for call and put options). We compute the relative distance of all observed IV from fitted IV and then we truncate the fitted curve to bins of moneyness with a width of 0.05(0.8, 0.85, ..., 1.2). After calculating the standard deviation for each moneyness bin, we discard any quote whose observed IV is one standard deviation apart from the fitted IV.

- Put-call parity filter: we remove any quote whose the put-call parity implied interest rate is more than one standard deviation away from the average among the peers. Peer group is defined as quotes with the same (date, time to maturity) pairs. We trim outliers in a similar way as with the IV filter. Specifically, we use the whole sample of distances of the put-call parity implied interest rates from the corresponding daily median implied interest rate to find the standard deviation of the corresponding distances.

Table A.6 records the number of observations at each filtering level that are removed. Before the filters, we have a total of 8,261,170 observations. Level 1 filters remove 10 observations. The zero volume criteria in the Level 2 filters eliminate the most observations (5,433,167) and level

3 filters eliminate 8.3% of observations. Figure A.6 plots the trading volume of first 4-month on 2016 and the number of trading volume is indicated by color. This clearly demonstrates that options are highly traded between 0.8 and 1.2 moneyness and options expiring after 1 year are seldom traded.

We consider a sample of option price data on 27 August 2008 to illustrate the filter process and check the no-arbitrage property of the processed result. The raw option data and calculated call price are shown in Table A.7. With forward price \$1465.40 the S&P 500 index price is \$1281.66, which means that the S&P 500 index dividend rate is higher than the risk-free interest rate. This is confirmed by the statistics of our full dataset. The average dividend rate is 2.53% and the interest rate is 1.17%. Figure A.7 displays the distribution of call prices after filters. We note that there are three maturities available on this trading date. Clearly, the options trading of deep OTM is very thin even for S&P 500 index options[15]. Although the dataset does not have a large coverage of call prices as high-frequency data, we argue that our dataset is efficient enough to cover the call price surface. Compare this with Chiang et al. (2016),who also follows Constantinides et al. (2013) and obtains a total 404,822 observations on S&P 500 options between 1996 and 2011, so our dataset is relatively large.

Furthermore, to compare our no arbitrage filter with Zhang and Xiang (2008), we plot the time value of call option price on 04/11/2003 in Figure A.9 and check whether the dataset shows no arbitrage. The time value of an option is defined as

$$c_{time}(k,\tau) = c(k,\tau) - e^{-rt} max(F_0 - K, 0) \tag{A.3}$$

$$p_{time}(k,\tau) = p(k,\tau) - e^{-rt} max(K - F_0, 0) \tag{A.4}$$

If the put call parity relationship holds, this means that time value of call price $c_{time}(k,\tau)$ and put price $p_{time}(k,\tau)$ are equal. Comparing Figure A.9 with Figure 2 in Zhang and Xiang (2008), we argue that our dataset is arbitrage free and the time values of call and put prices almost coincide. The difference in peak time value of call price in Zhang and Xiang (2008) is significantly higher than ours. That said, our dataset contains less noise than Zhang and Xiang (2008).

[15]The deep of OTM option means the strike price is far away from current index price. Consequently, buying a deep OTM option is expecting that a extreme increase of index price.

**Table A.6: Number of Observations**

| | | Deleted | Remaining |
|---|---|---|---|
| Starting | Calls | | 4,130,624 |
| | Puts | | 4,130,546 |
| | All | | 8,261,170 |
| Accuracy Filters | Identical Terms | 10 | |
| | All | | 8,261,160 |
| Liquidity Filters | Zero Bid /Ask | 708,202 | |
| | Zero Volume | 5,433,167 | |
| | Days to Expiration Boundaries | 262,388 | |
| | Moneyness Boundaries | 254,545 | |
| | IV Boudaries | 417,832 | |
| | All | | 1,185,026 |
| NA validation | Negative Implied Interest Rate | 289,542 | |
| | Negative Time Value | 232 | |
| | IV Outlier | 173,725 | |
| | Implied Interest Rate Outlier | 222,749 | |
| | All | | 498,778 |
| OTM Options | | | 390,320 |

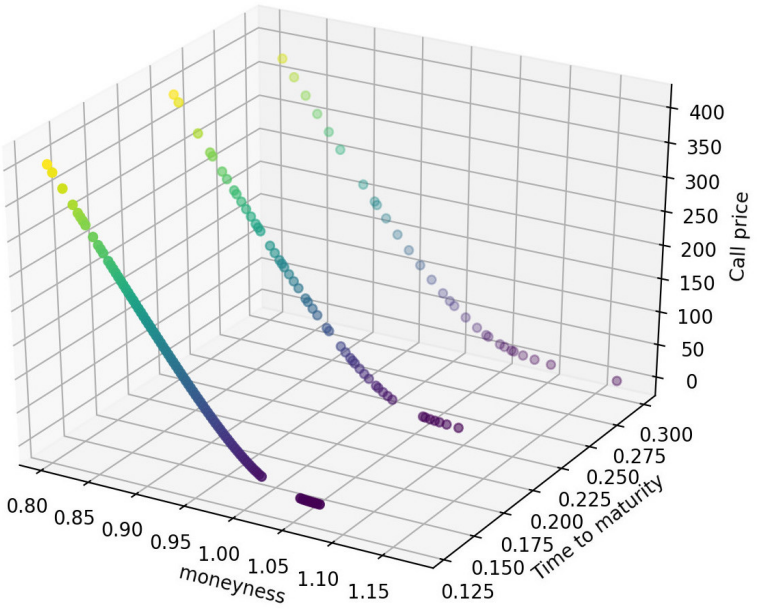
**Note**: Table A.6 presents the number of observations after each filter. The sample period is from 01/01/2000 to 04/30/2016. The moneyness is defined as the ratio of strike price to index price. IV is implied volatility, NA is no arbitrage and OTM is out the money. Implied interest rate is calculated based on linear regression. The time value of option is defined in Equation(A.3) and (A.4).

**Table A.7: Option Data Panel on August 27, 2008**

| Options | Strike | Bid | Ask | Volume | Call Option Price |
|---|---|---|---|---|---|
| | 1435 | 20 | 22.2 | 3177 | 51.40 |
| | 1440 | 20.9 | 23.5 | 7168 | 47.52 |
| | 1445 | 22.4 | 25 | 5322 | 44.03 |
| Put | 1450 | 24.2 | 25.3 | 25947 | 40.10 |
| | 1455 | 25.6 | 28.2 | 2594 | 37.26 |
| | 1460 | 27.5 | 29.9 | 3339 | 34.08 |
| | 1465 | 29.2 | 31.8 | 10440 | 30.89 |
| ATM | 1465.40 | | | | |
| | 1470 | 26.5 | 29.1 | 142 | 27.80 |
| | 1475 | 25 | 25.8 | 4931 | 25.40 |
| | 1480 | 22 | 23.5 | 2574 | 22.75 |
| Call | 1485 | 18.5 | 20.9 | 355 | 19.70 |
| | 1490 | 16.1 | 18.5 | 888 | 17.30 |
| | 1495 | 13.9 | 16.3 | 265 | 15.10 |
| | 1500 | 12.4 | 13.5 | 29721 | 12.95 |

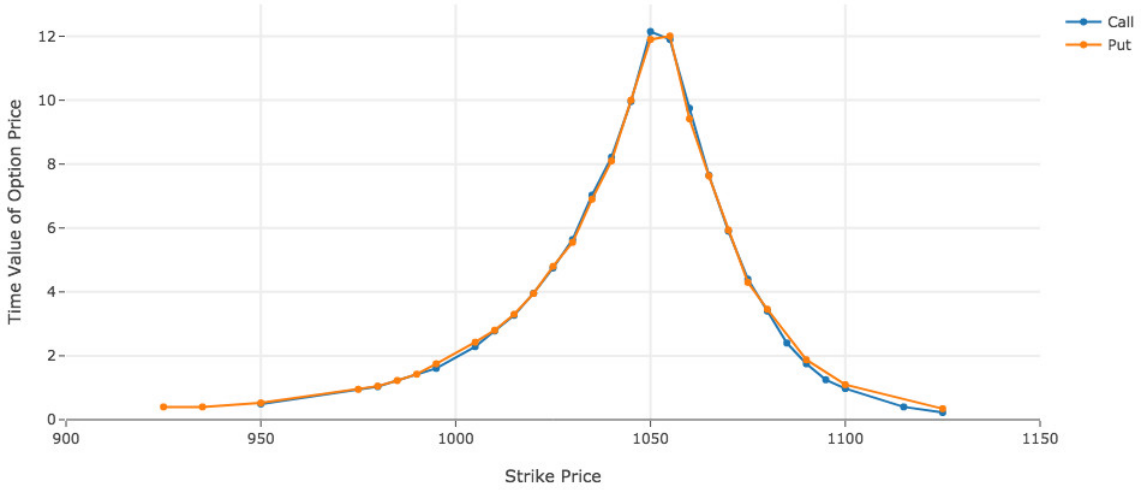
**Note**: Table A.7 shows the market data on 27/08/2007. The days to maturity is 24. The computed ATM forward price is \$1465.40 and S&P 500 index price is \$1281.66. To give reader a general idea, this table only presents a part of data.

**Figure A.7: Distribution of Call Option Price after Filters on August 27, 2008**




**Note**: This figure shows the S&P 500 index call option price across different strike and time to maturity on 27/08/2008. The S&P 500 index price is \$1281.66.

**Figure A.8: Time Value of S&P 500 Option Price on November 04, 2003**




**Note**: This figure shows the time value of S&P 500 index call and put options on 04/11/2003. The days to maturity is 17 and the S&P 500 index price is $1053.25.

## AppendixB. Change measure of Call Price

To compare our $L_1$-SVM estimator with other models, we investigate the estimated call price problem under the forward measure (from $\mathbb{Q}$ to $\mathbb{Q}^{\mathbb{T}}$). Define $F(t,T)$ as the forward measure and $F(T,T)=1$.

$$E^{\mathbb{Q}}[V(T)|\mathcal{F}] = E^{\mathbb{Q}^{\mathbb{T}}}[V(T)\frac{M(t)P(T,T)}{M(T)P(t,T)}|\mathcal{F}] \tag{B.1}$$

$$\frac{d\mathbb{Q}^{\mathbb{T}}}{d\mathbb{Q}} = \frac{e^{\int_t^T r(t)dt}}{F(t,T)} \tag{B.2}$$

Previous studies make different assumptions regarding interest rate and dividends. Some studies allow constant interest rate and some for a deterministic rate (see Table 4 for summary). To compare existing estimators with the proposed estimator, we normalize the call price to eliminate the influence of interest rate and dividend. Recall Merton's (1973) model, which includes dividends in the Black-Scholes framework. We define forward price $F(t,T) = se^{(r-\delta)\tau}$ and simplify the $e^{\int_t^T r(t)dt}$ as $e^{rt}$.

$$C(K,\tau) = se^{-\delta\tau}N(d_1) - Ke^{-r\tau}N(d_2) \tag{B.3}$$

$$\frac{C(K,\tau)e^{rt}}{se^{(r-\delta)\tau}} = \frac{se^{-\delta\tau}N(d_1)e^{rt}}{se^{(r-\delta)\tau}} - \frac{e^{rt}Ke^{-r\tau}N(d_2)}{se^{(r-\delta)\tau}} \tag{B.4}$$

$$\frac{C(K,\tau)}{se^{-\delta\tau}} = N(d_1) - \frac{Ke^{-r\tau}N(d_2)}{se^{-\delta\tau}} \tag{B.5}$$

If we define the forward moneyness $k = \frac{K}{F(t,T)} = \frac{K}{se^{(r-\delta)\tau}}$ , then the Black-Scholes-Merton formula can transform to

$$\frac{C(K,\tau)e^{rt}}{F(t,T)} = N(d_1) - kN(d_2) \tag{B.6}$$

$$d_1 = \frac{ln(\frac{s}{K}) + (r_{t,\tau} - \delta_\tau + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}} = \frac{ln(\frac{1}{k}) + \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}} \tag{B.7}$$

$$d_2 = d_1 - \sigma\sqrt{\tau} \tag{B.8}$$

Compare Equation (B.3) with Equation (B.6). Equation (B.6) is equivalent to a European call option price with underlying is 1, interest rate $r$ and dividend yield $\delta$ equal to 0. Hence, by transforming the estimating framework to forward measure, we eliminate the influence the interest rate and dividend yield.

## AppendixC. Proof of Roper (2010) under the forward measure

There are two ways to restrict the arbitrage free call price. One is from the state price density perspective and the other is from the option strategies perspective. In this section, we show how to get similar no arbitrage conditions to those of Roper (2010) under the forward measure. First, recall the no arbitrage conditions under the risk neutral measure $\mathbb{Q}$. The call price is the numerical integration of the payoff and risk neutral density (RND).
Define $k = \frac{K}{F(t,T)}$

$$C(S_t,k,\tau,r,\delta) = E_t^{Q^T}[max(\frac{S_T}{F(t,T)} - \frac{K}{F(t,T)},0)] = \int_0^\infty \max(\frac{S_T}{F(t,T)} - k,0)q^{Q^T}(S_T)dS_T \tag{C.1}$$

Take the first order derivative

$$\frac{\partial C(k,\tau)}{\partial k} = -\int_0^\infty q^{Q^T}(S_T)dS_T = (1 - F(x)) \tag{C.2}$$

Where $F(x)$ is cumulative distribution function of the transition probability $q^{Q^T}$ under forward measure

$$\Rightarrow \frac{\partial C(k,\tau)}{\partial k} = (F(x) - 1) \tag{C.3}$$

As $F(x)$ is always greater than zero:

$$\Rightarrow F(x) \geq 0$$

$$\Rightarrow \frac{\partial C(k,\tau)}{\partial k} \geq -1 \tag{C.4}$$

and the transition probability integrates to one:

$$\int_0^\infty q^{Q^T}(S_T) dS_T = 1 \tag{C.5}$$

$$\Rightarrow q^{Q^T}(S_T) > 0$$

$$\Rightarrow \frac{\partial C(k,\tau)}{\partial k} \leq 0$$

Therefore

$$-1 \leq \frac{\partial C(k,\tau)}{\partial k} \leq 0 \tag{C.6}$$

Further differentiate Equation (C.4),

$$\frac{\partial C(k,\tau)}{\partial k} = q^{Q^T}(S^T) \geq 0 \tag{C.7}$$

Since $C(S_t, k, \tau, r, \delta) = E_t^{Q^T}[max(\frac{S_T}{F(t,T)} - \frac{K}{F(t,T)}, 0)]$, by Jensen's inequality(Jensen (1906)), the max function is convex, thus

$$C(S_t, k, \tau, r, \delta) \geq max(E_t^{Q^T}[(\frac{S_T}{F(t,T)} - k, 0)] = max(1 - k, 0) \tag{C.8}$$

This result can also be obtained by setting $D(T)F(T) = 1$ and $S(T) = 1$ in Equation(2.27) and (2.32). When $k = 0$, $C(S_t, k = 0, \tau, r, \delta) = E_t^{Q^T}[\frac{S_T}{F(t,T)}] = 1$. Intuitively, following Equa-

tion(4.19), the price bounds under the forward measure is

$$0 < C(k, \tau) < 1 \tag{C.9}$$

Similarly, we can prove (C3) that as $k \to \infty$, the option becomes worthless. From a practical point of view, an investor would not buy this option because it is impossible to exercise it. C(5) is easy to prove via the convexity and monotonicity of call option price.

Following Aït-Sahalia and Duarte (2003) and Fengler and Hin (2015), we argue that using the homogeneity assumption, C(1),C(2) and C(4) are sufficient to restrict an arbitrage free call price surface. Therefore in Equation (30), we only consider incorporating C(1),C(2) and C(4) in the machine learning framework $L_1 - SVM$.